Explainable Autonomy through Natural Language

Francisco Javier Chiyah Garcia

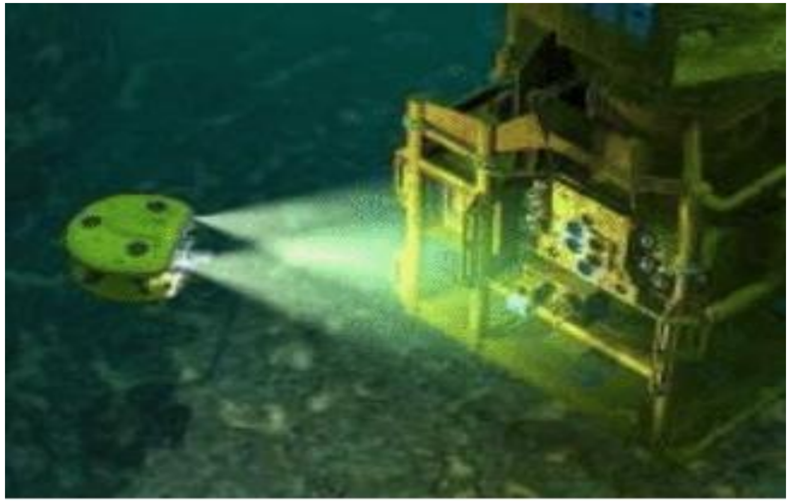
Heriot-Watt University, Edinburgh, UK

GI-Dagstuhl Seminar 2019 – Explainable Software for Cyber-Physical Systems

About me

- 5th Year student of MEng in Software Engineering.
- Worked for 6 months at SeeByte (software for underwater vehicles and sensors).
- Main contribution: MIRIAM, a multimodal interface for autonomous underwater vehicles.
- Areas: explainability, NLP, NLG, autonomy, augmented-reality…
- Human-Robot Interaction centred.

HERIOT WATT UNIVERSITY



Robots and Autonomous Systems

- Increasingly being operated remotely, particularly in hazardous environments (Hastie et al., 2018).
- These can instil less trust (Bainbridge et al., 2008).
- Thus, the interface between operator and autonomous systems is key (Robb et al., 2018).

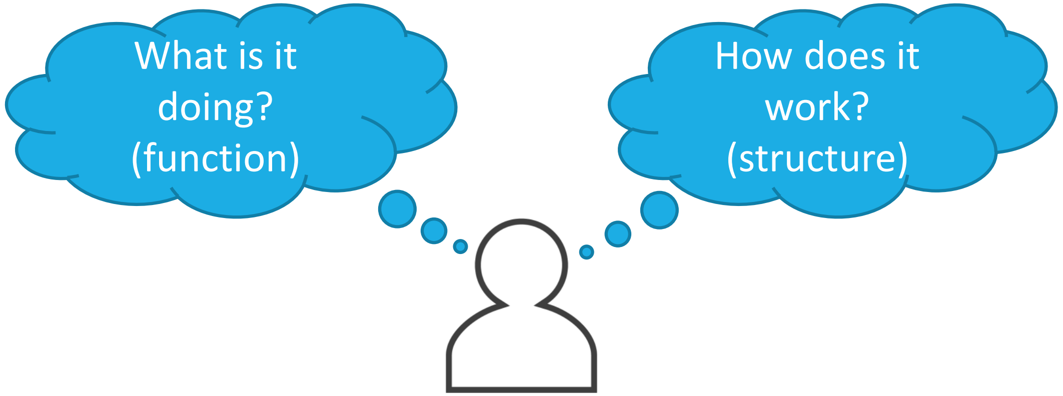
Transparency

- Robots and autonomous systems are hard to understand for non-experts.
- This lack of transparency of how a robot behaves is reflected in decreased trust and understanding.
- Decreased trust and understanding have negative effects on human-machine cooperation.
- Transparent systems are able to provide explanations.

Trust in Autonomous Systems

Mental Models and Explanations 1

- Mental models strongly impact how and whether systems are used.
- Explanations contribute to building accurate mental models of a system.
- Improving the user’s mental model can provide increased confidence and performance (Le Bras et al., 2018).
- According to (Gregor and Benbasat, 1999; Kulesza et al., 2013), “users will not expend effort to find explanations unless the expected benefit outweighs the mental effort”.

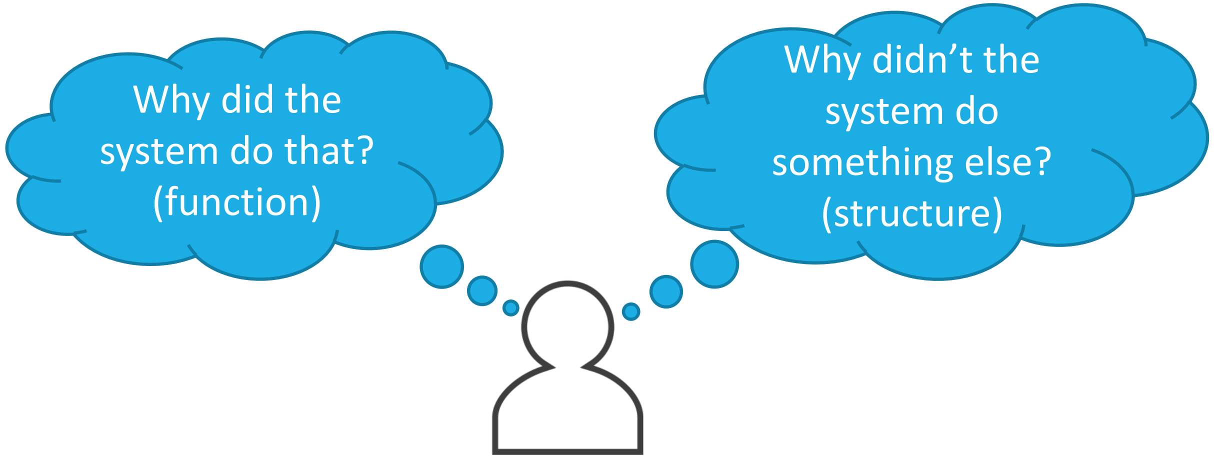


What is it doing? (function)

How does it work? (structure)

Mental Models and Explanations 2

- Lim et al. (2009) showed that:
  - explaining "*why*" a system behaved in a certain way increased understanding and trust
  - "*why not*" explanations only increased understanding
- Thus both are important regarding the user's mental model.

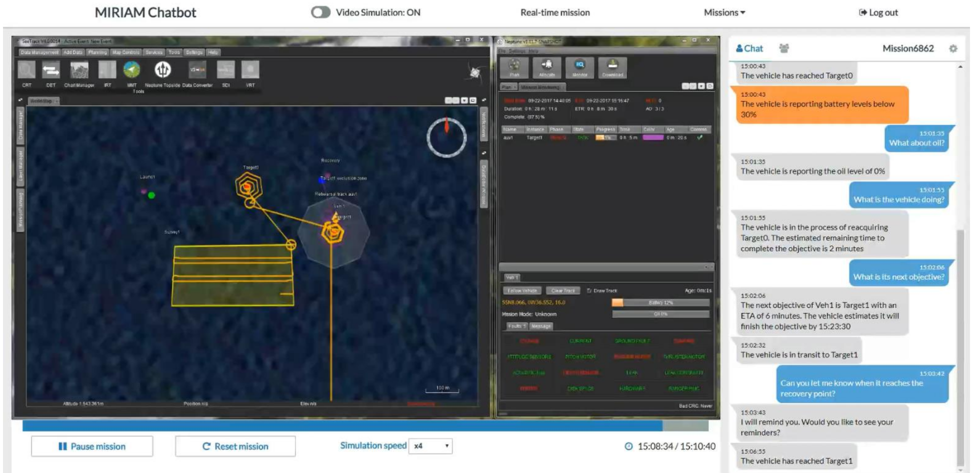


Why did the system do that? (function)

Why didn't the system do something else? (structure)

MIRIAM: The Multimodal Interface 1

- MIRIAM allows for “on-demand” queries for status and explanations of behaviour.
- Increases the user’s situation awareness.
- Requires little training.



15:00:43
The vehicle has reached Target0

15:00:43
The vehicle is reporting battery levels below 30%

15:01:35
What about oil?

15:01:35
The vehicle is reporting the oil level of 0%

15:01:55
What is the vehicle doing?

15:01:55
The vehicle is in the process of reacquiring Target0. The estimated remaining time to complete the objective is 2 minutes

15:02:06
What is its next objective?

15:02:06
The next objective of Veh1 is Target1 with an ETA of 6 minutes. The vehicle estimates it will finish the objective by 15:23:30

15:02:32
The vehicle is in transit to Target1

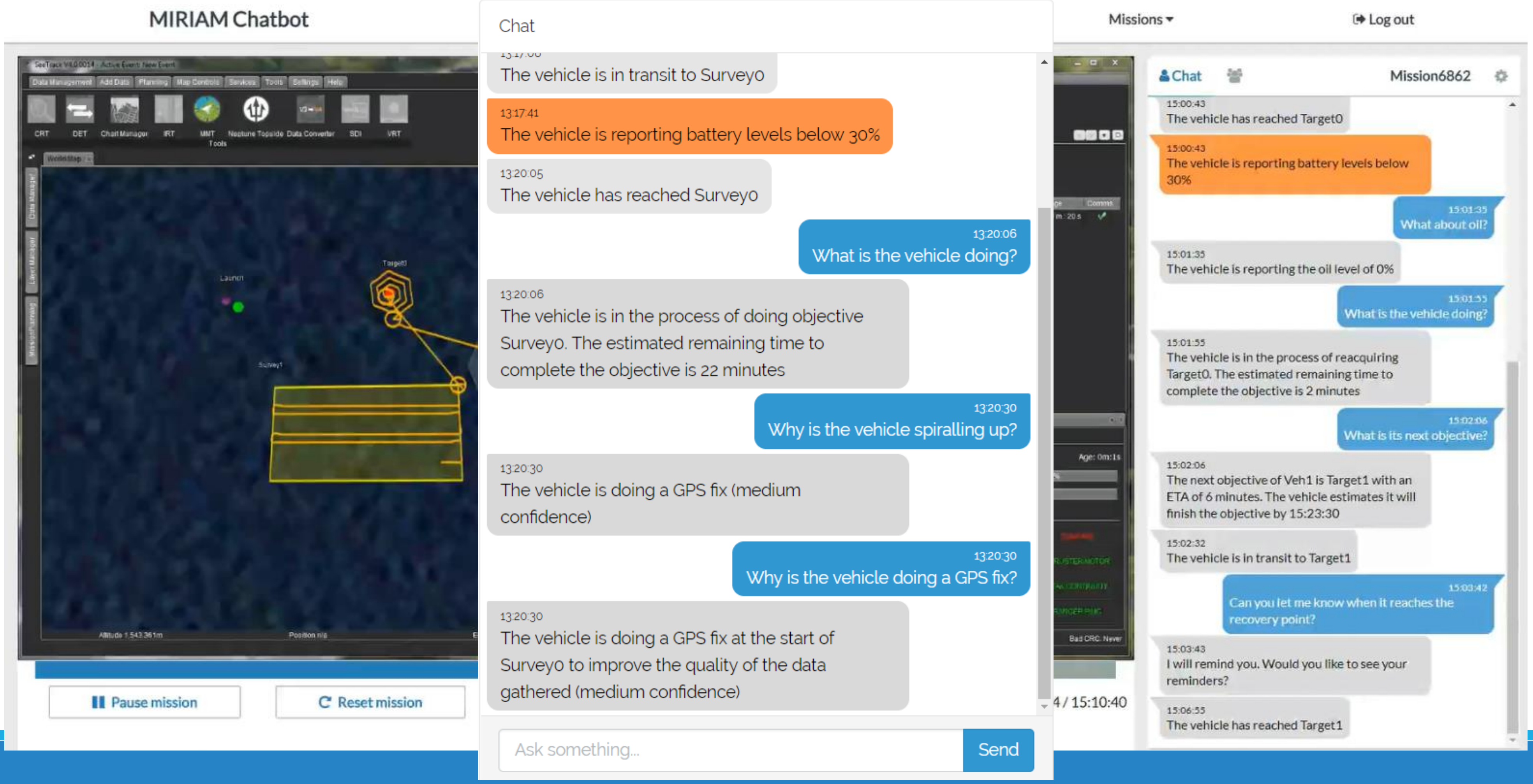
15:03:42
Can you let me know when it reaches the recovery point?

15:03:43
I will remind you. Would you like to see your reminders?

15:06:55
The vehicle has reached Target1

Hastie, Helen; Chiyah Garcia, Francisco J.; Robb, David A.; Patron, Pedro; Laskov, Atanas: MIRIAM: A Multimodal Chat-Based Interface for Autonomous Systems. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI’17. ACM, Glasgow, UK, pp. 495–496, 2017.

MIRIAM: The Multimodal Interface 2



Chat

13:17:00
The vehicle is in transit to Survey0

13:17:41
The vehicle is reporting battery levels below 30%

13:20:05
The vehicle has reached Survey0

13:20:06
What is the vehicle doing?

13:20:06
The vehicle is in the process of doing objective Survey0. The estimated remaining time to complete the objective is 22 minutes

13:20:30
Why is the vehicle spiralling up?

13:20:30
The vehicle is doing a GPS fix (medium confidence)

13:20:30
Why is the vehicle doing a GPS fix?

13:20:30
The vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (medium confidence)

Chat — Mission6862

15:00:43
The vehicle has reached Target0

15:00:43
The vehicle is reporting battery levels below 30%

15:01:35
What about oil?

15:01:35
The vehicle is reporting the oil level of 0%

15:01:55
What is the vehicle doing?

15:01:55
The vehicle is in the process of reacquiring Target0. The estimated remaining time to complete the objective is 2 minutes

15:02:06
What is its next objective?

15:02:06
The next objective of Veh1 is Target1 with an ETA of 6 minutes. The vehicle estimates it will finish the objective by 15:23:30

15:02:32
The vehicle is in transit to Target1

15:03:42
Can you let me know when it reaches the recovery point?

15:03:43
I will remind you. Would you like to see your reminders?

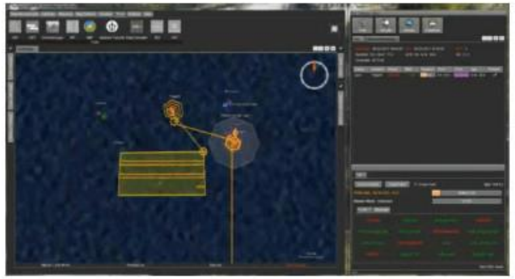
15:06:55
The vehicle has reached Target1

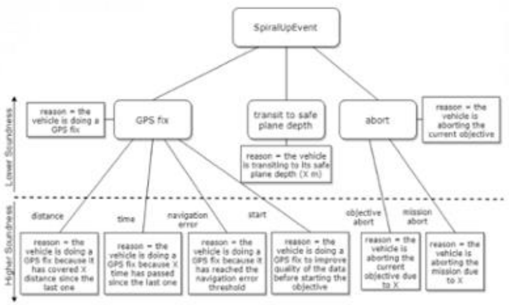
seebyte

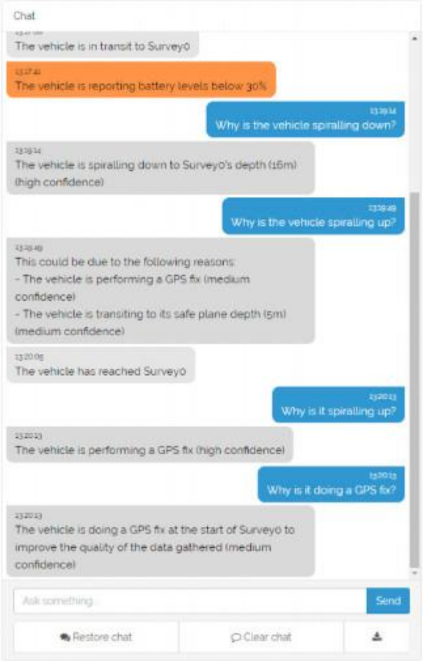
Explainability

- The conversational agent can:
  - Give information about *what* is happening (*function*)
    e.g. “What is the vehicle doing?”, “What is the battery level of the vehicle?”
  - Explain *why* the vehicles are doing (or did) something (*function*)
    e.g. “Why is the vehicle coming to the surface?”
  - Explain “*why not”* the vehicles did not do an *expected action* (*structure*)
    e.g. “Why is the vehicle not going to Area 1?”

“Why” and “Why not” Explanations




SpiralUpEvent
GPS fix
transit to safe plane depth
abort
reason = the vehicle is doing a GPS fix
reason = the vehicle is transiting to its safe plane depth (X m)
reason = the vehicle is aborting the current objective
Lower Soundness
Higher Soundness
distance
time
navigation error
start
objective abort
mission abort
reason = the vehicle is doing a GPS fix because it has covered X distance since the last one
reason = the vehicle is doing a GPS fix because X time has passed since the last one
reason = the vehicle is doing a GPS fix because it has reached the navigation error threshold
reason = the vehicle is doing a GPS fix to improve quality of the data before starting the objective
reason = the vehicle is aborting the current objective due to X
reason = the vehicle is aborting the mission due to X



Chat

The vehicle is in transit to Survey0

13:17:41
The vehicle is reporting battery levels below 30%

13:19:14
Why is the vehicle spiralling down?

13:19:14
The vehicle is spiralling down to Survey0's depth (16m) (high confidence)

13:19:49
Why is the vehicle spiralling up?

13:19:49
This could be due to the following reasons:
- The vehicle is performing a GPS fix (medium confidence)
- The vehicle is transiting to its safe plane depth (5m) (medium confidence)

13:20:05
The vehicle has reached Survey0

13:20:13
Why is it spiralling up?

13:20:13
The vehicle is performing a GPS fix (high confidence)

13:20:13
Why is it doing a GPS fix?

13:20:13
The vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (medium confidence)

Expert

Mental models align easily

User

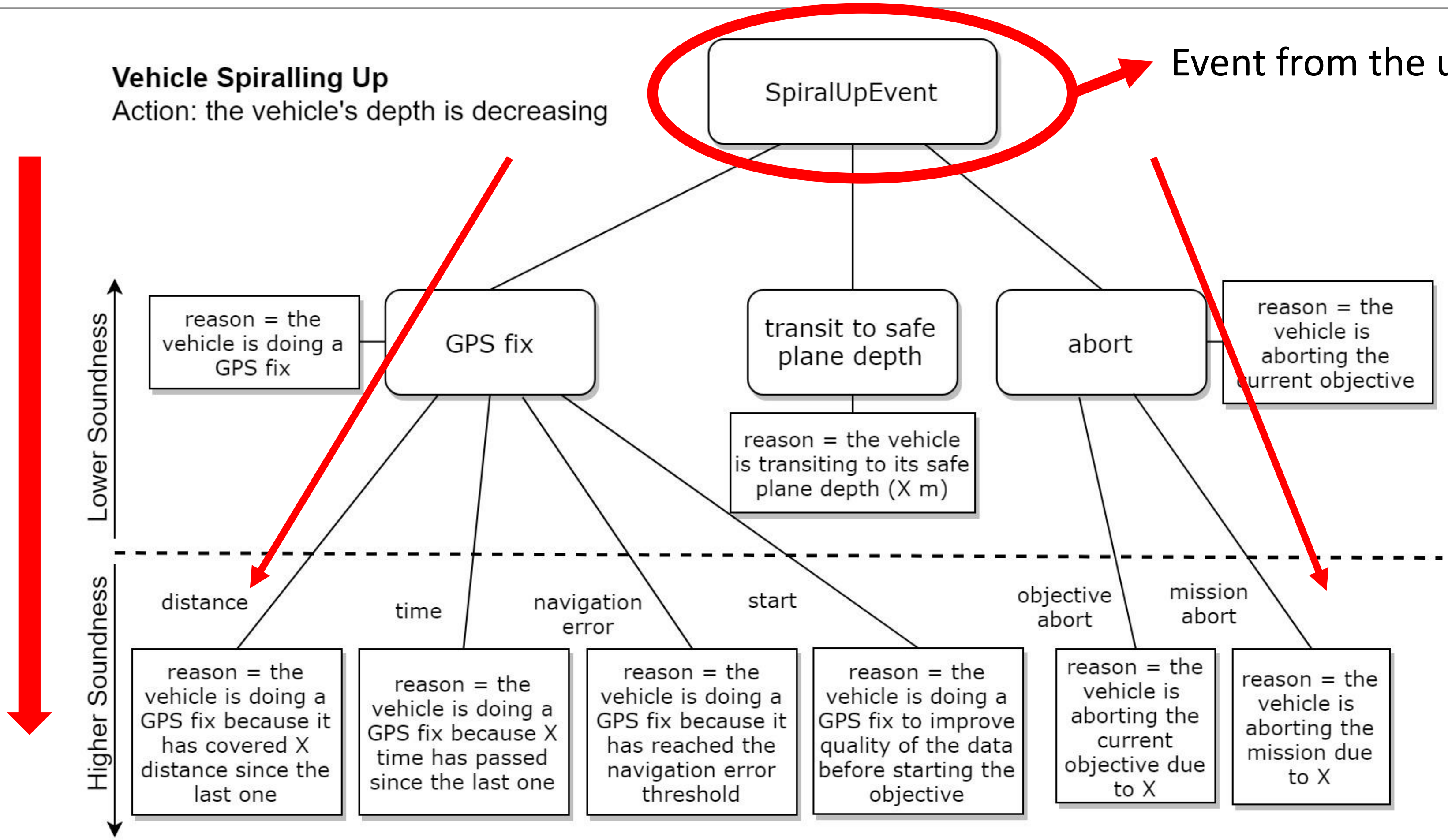
Generation Method 1

- ‘Speak-aloud’ method whereby an expert provides rationalisation of the autonomous behaviours.
- Derive a model of autonomy.
- Data received from the vehicles is used to steadily build a knowledge base.

Two autonomous underwater vehicles.

Model of Autonomy

Traversing down provides the trace for “why” or “why not” explanations



Vehicle Spiralling Up
Action: the vehicle's depth is decreasing

SpiralUpEvent

Event from the user’s perspective

GPS fix

reason = the vehicle is doing a GPS fix

transit to safe plane depth

reason = the vehicle is transiting to its safe plane depth (X m)

abort

reason = the vehicle is aborting the current objective

reason = the vehicle is doing a GPS fix because it has covered X distance since the last one

reason = the vehicle is doing a GPS fix because X time has passed since the last one

reason = the vehicle is doing a GPS fix because it has reached the navigation error threshold

reason = the vehicle is doing a GPS fix to improve quality of the data before starting the objective

reason = the vehicle is aborting the current objective due to X

reason = the vehicle is aborting the mission due to X

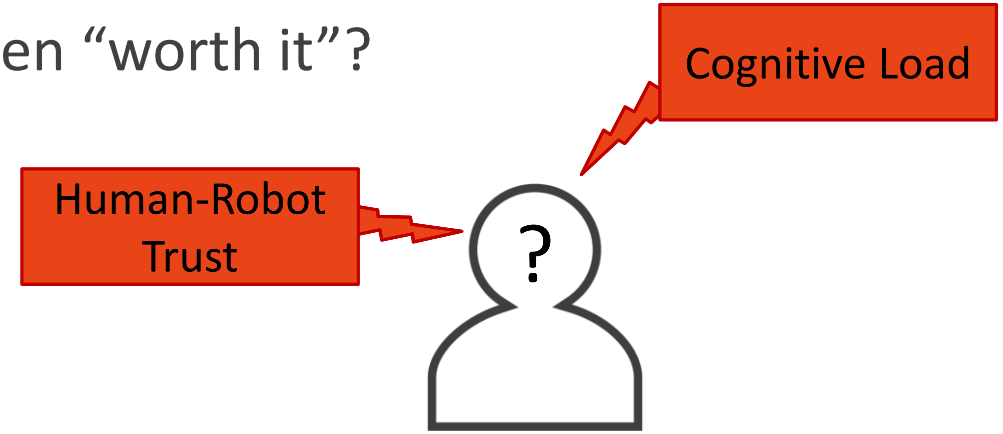
Generation Method 2

- Explanations are generated on-demand from a dynamic database that captures context.
- Template-based NLG.
- Explanations come with a confidence value.

- Example explanation:
  - **User:** Why is the vehicle coming to the surface?
  - **System:** The vehicle is transiting to its safe plane depth (medium confidence).

Explanation Effects

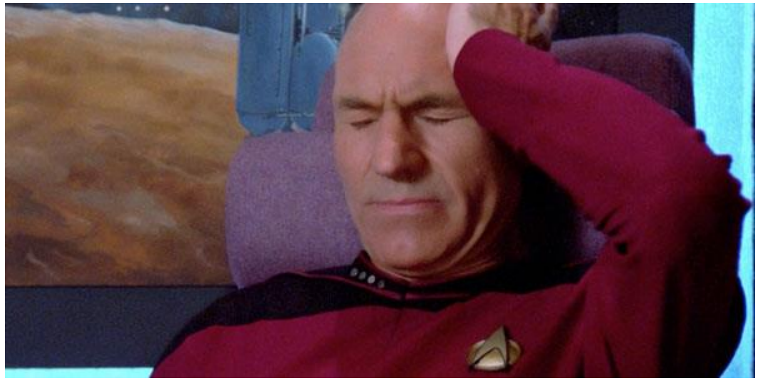
- Investigated the effects of explanations on the user through a study.
- What is the best way to give an explanation?
- “*What*” and “*how*” to say it are both important.
- **Level of detail of an explanation** vs **number of autonomy model reasons** (soundness vs completeness)
- Are they even “worth it”?



Human-Robot Trust

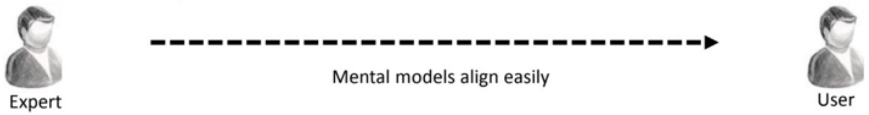
Cognitive Load

?

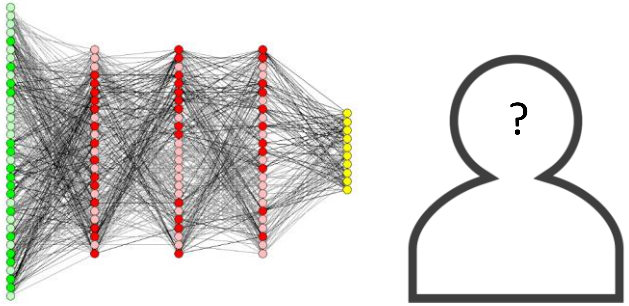


Method Insights

- **Advantages:**
  - Expert knowledge can be transferred easily
  - High-level abstraction
  - User-centred
  - On-demand



- **Disadvantages:**
  - Manual process ('speak-aloud')
  - Scalability
  - ML systems may prove hard for an expert to explain



Future Work

- Expand what the conversational agent can understand and process
  - ➢ Could we do this automatically?

- Generalisation of the agent
  - ➢ Could the agent be useful in other domains/systems?

- Handle uncertainty better
  - ➢ What are the best ways to handle it?

Summary

- Understanding *what* a system does and *how* it works is important.
- Transparent systems are able to provide explanations.
- Different types of explanations and effects: “why”, “why not”.
- Users won’t read explanations if they don’t believe it is worth it.
- A conversational agent that gives on-demand information.

ES4CPS

- **What is an ES4CPS problem, and/or what is an ES4CPS solution, that I am interested in?**
  - What makes a system explainable? Can we achieve a formal definition?
  - Conversational agents as an intuitive way of explaining a system on-demand.
- **What is the ES4CPS-related expertise that I can contribute to solving this problem?**
  - Human-Robot Interaction.
  - Experience with explanations (why, why not) and their effects.
- **What external expertise do I need (possibly from the other participants) in order to work on the problem/solution?**
  - Distinct concepts of explainability, discuss what it aims to achieve.
  - Expertise with other explainable systems.

Acknowledgements

HERIOT WATT UNIVERSITY

- Prof. Helen Hastie
- Dr. David A. Robb

seebyte

- Dr. Pedro Patron
- Atanas Laskov

References

- Hastie, Helen; Lohan, Katrin Solveig; Chantler, Mike J.; Robb, David A.; Ramamoorthy, Subramanian; Petrick, Ron; Vijayakumar, Sethu; Lane, David: The ORCA Hub: Explainable Offshore Robotics through Intelligent Interfaces. In: Proceedings of Explainable Robotic Systems Workshop, HRI'18. Chicago, IL, USA, 2018.

- Kulesza, Todd; Stumpf, Simone; Burnett, Margaret; Yang, Sherry; Kwan, Irwin; Wong, Weng-Keen: Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing. San Jose, CA, USA, pp. 3–10, Sept 2013.

- Le Bras, Pierre; Robb, David A.; Methven, Thomas S.; Padilla, Stefano; Chantler, Mike J.: Improving User Confidence in Concept Maps: Exploring Data Driven Explanations. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–13, 2018.

- Lim, Brian Y.; Dey, Anind K.; Avrahami, Daniel: Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '09, pp. 2119–2129, 2009.

- Robb, David A.; Chiyah Garcia, Francisco J.; Laskov, Atanas; Liu, Xingkun; Patron, Pedro; Hastie, Helen: Keep Me in the Loop: Increasing Operator Situation Awareness through a Conversational Multimodal Interface. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. ICMI'18, ACM, Boulder, Colorado, USA, 2018.

HERIOT WATT UNIVERSITY

Thank you for your attention

QUESTIONS?

Explainable Autonomy through Natural Language
F. J. Chiyah Garcia
fjc3@hw.ac.uk